

Initiation à l'ACM : analyse des correspondances multiples


J.R. Lobry & A.B. Dufour

Un jeu de données très simple qui peut être analysé aussi bien en ACP qu'en ACM est utilisé pour une première approche introductive à l'ACM.

Table des matières

1	Les données	2
2	Analyse univariée	2
2.1	Les 8 premières variables	2
2.1.1	altit	2
2.1.2	deniv	3
2.1.3	cloiso	3
2.1.4	domain	4
2.1.5	boise	4
2.1.6	hetra	4
2.1.7	favor	5
2.1.8	inexp	5
2.2	Les 2 dernières variables	6
2.2.1	citat	6
2.2.2	depart	6
3	Une ACP pour commencer	7
4	Une ACM pour continuer	9
4.1	Introduction	9
4.2	Mise en oeuvre	10
4.3	Une différence importante en l'ACP et l'ACM	13
	Références	15

1 Les données

Les données utilisées sont disponibles dans le paquet  `ade4`. Elles ont été établies lors de la rédaction d'un ouvrage [1] sur l'éradication de l'ours brun dans les Alpes françaises.

```
library(ade4)
data(ours)
dim(ours)
[1] 38 10
names(ours)
[1] "altit" "deniv" "cloiso" "domain" "boise" "hetra" "favor" "inexp" "citat"
[10] "depart"
summary(ours)
altit deniv cloiso domain boise hetra favor inexp citat depart
1: 8 1:13 1:12 1: 9 1:10 1:19 1:15 1:20 1:22 AHP:5
2:17 2:14 2: 4 2:13 2:15 2: 5 2:12 2:10 2: 7 AM :4
3:13 3:11 3:22 3:16 3:13 3:14 3:11 3: 8 3: 4 D :5
4: 5 HP :8
HS :4
I :5
S :7
```

2 Analyse univariée

Nous allons commencer par faire l'examen univarié de ces données. Les individus sont ici 38 régions définies par l'Inventaire National Forestier et caractérisées par 10 variables. Les 8 premières variables décrivent les caractéristiques environnementales liées à la présence de l'ours ; les deux dernières représentent des informations complémentaires.

2.1 Les 8 premières variables

Les 8 premières variables ont la même logique et peuvent être considérées comme des variables qualitatives ordonnées ou comme des indices quantitatifs, les notes 1, 2 et 3 codant dans l'ordre une situation *a priori* de plus en plus favorable à l'ours brun (grands espaces forestiers, connexes, d'accès difficile).

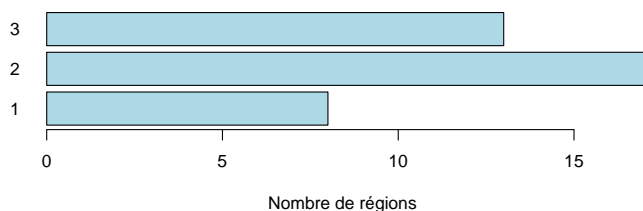
2.1.1 altit

La variable donne l'importance de la tranche altitudinale habitée par l'ours (800-2000 m) sous forme d'un facteur à 3 modalités :

1. moins 50 % de la surface se situe entre 800 et 2000 m
2. entre 50 et 70 % de la surface se situe entre 800 et 2000 m
3. plus de 70 % de la surface se situe entre 800 et 2000 m

Représenter graphiquement les effectifs des modalités de cette variable qualitative :

La variable altit

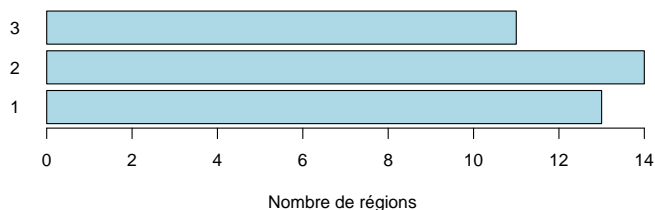


2.1.2 deniv

La variable donne l'importance du dénivelé moyen par carré de 50km^2 sous forme d'un facteur à 3 modalités :

1. moins de 700 m
2. entre 700 et 900 m
3. plus de 900 m

La variable deniv

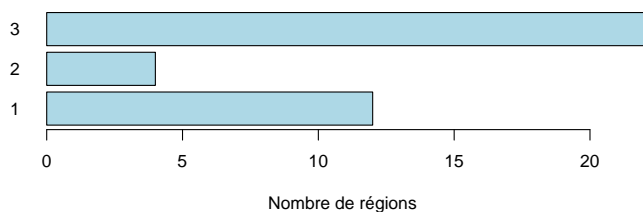


2.1.3 cloiso

La variable décrit le cloisonnement du secteur sous la forme d'un facteur à 3 modalités :

1. une grande vallée ou une crête isole au moins un quart du massif
2. une grande vallée ou une crête isole moins d'un quart du massif
3. le massif ne présente pas de coupure

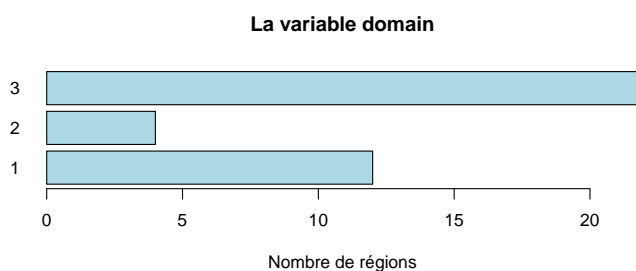
La variable cloiso



2.1.4 domain

La variable donne l'importance du domaine forestier en contact avec la région sous la forme d'un facteur à 3 modalités :

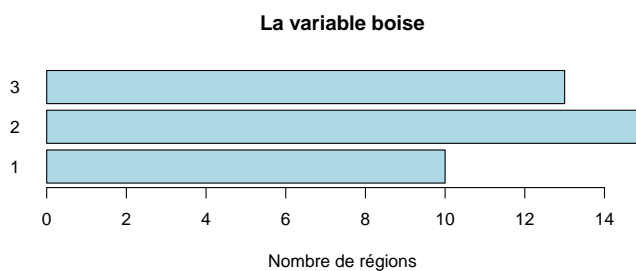
1. moins de 400 km²
2. entre 400 et 1000 km²
3. plus de 1000 km²



2.1.5 boise

La variable donne le taux (pourcentage de surface de la région) de boisement sous la forme d'un facteur à 3 modalités :

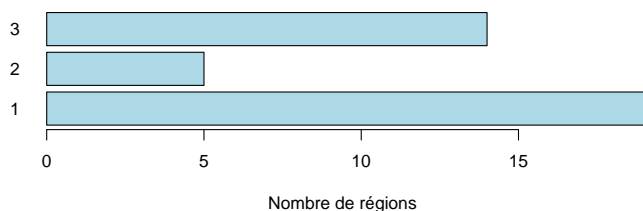
1. moins de 30 %
2. entre 30 et 50 %
3. plus de 50 %



2.1.6 hetra

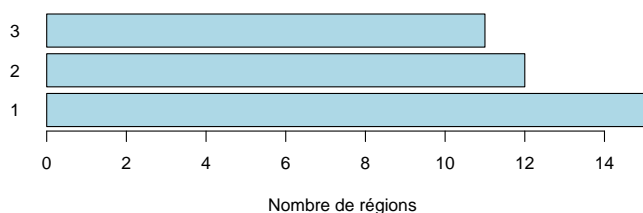
La variable indique l'importance des hêtraies et forêts mixtes dans la région sous forme d'un facteur à 3 modalités :

1. moins de 5%
2. entre 5 et 10%
3. plus de 10%

La variable hetra**2.1.7 favor**

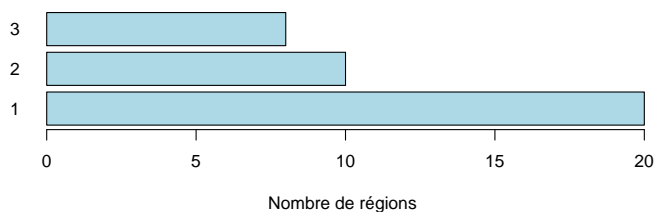
La variable renseigne l'importance (pourcentage de surface de la région) des forêts favorables à l'ours (hêtraies, forêts mixtes, sapinières et pessières) sous forme d'un facteur à 3 modalités :

1. moins de 5 %
2. entre 5 et 10 %
3. plus de 10 % du massif

La variable favor**2.1.8 inexp**

La variable renseigne l'importance (pourcentage de surface de la région) des forêts inexploitées sous forme d'un facteur à 3 modalités :

1. moins de 4 %
2. entre 4 et 8 %
3. plus de 8 % du massif

La variable inexp

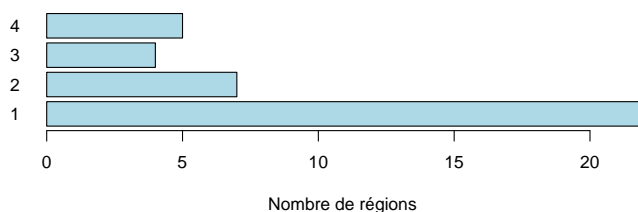
2.2 Les 2 dernières variables

2.2.1 citat

La variable donne l'information sur la date de disparition de l'Ours sous la forme d'une variable qualitative ordonnée à 4 modalités :

1. aucune citation de l'espèce depuis 1840
2. 1 à 3 citations avant 1900 et aucune après
3. 4 citations et plus avant 1900 et aucune après
4. 1 citation ou plus entre 1900 et 1940

La variable citat



2.2.2 depart

La variable indique le département qui contient la région sous la forme d'une variable qualitative à 7 modalités :

AHP Alpes-de-Haute-Provence

AM Alpes-Maritimes

D Drôme

HP Hautes-Alpes

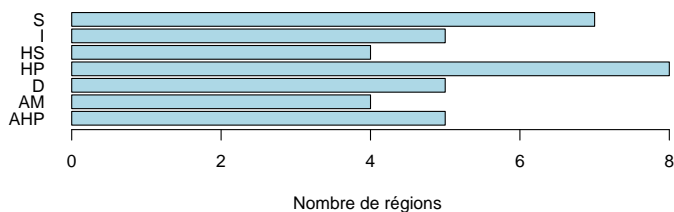
HS Haute-Savoie

I Isère

S Savoie



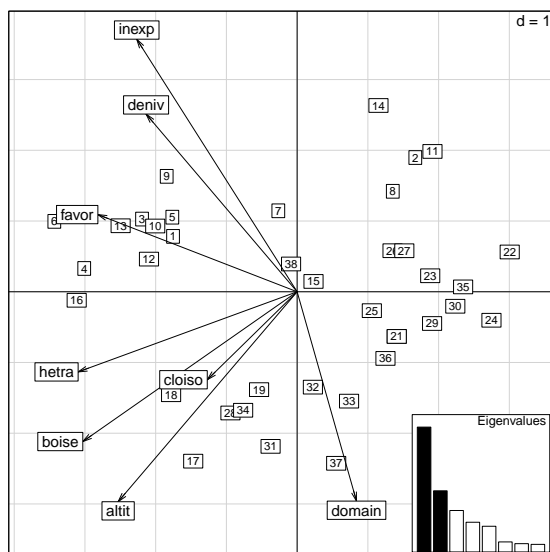
La variable depart



3 Une ACP pour commencer

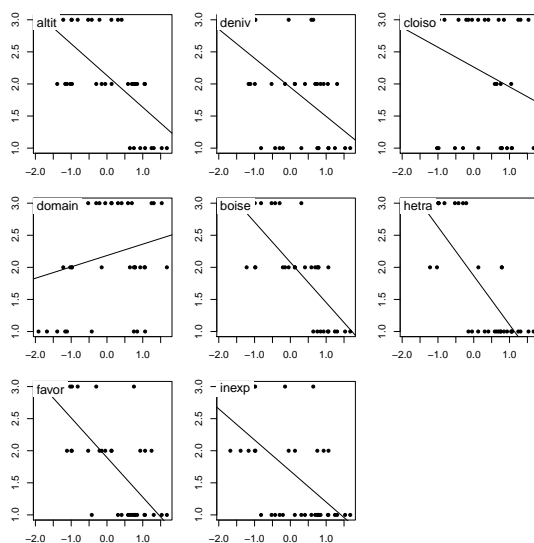
Les huit premières variables étant des variables qualitatives ordonnées associées à une description de l'environnement, on peut les transformer en variables semi-quantitatives sans perte de sens et faire une ACP pour résumer l'information.

```
qtt <- apply(ours[, 1:8], 2, as.numeric)
acp <- dudi.pca(qtt, scan = FALSE)
scatter(acp, posieig = "bottomright")
```



Le résultat est sans surprise, les variables sont globalement toutes corrélées positivement entre elles, ce qui est dans leur logique même. On note aussi que la variable **domain** est particulière puisque que c'est la seule à être de signe opposé aux autres sur le premier facteur. Pour aider à l'interprétation, on considère le graphe canonique qui confronte les coordonnées sur le premier facteur aux valeurs originelles :

```
score(acp)
```

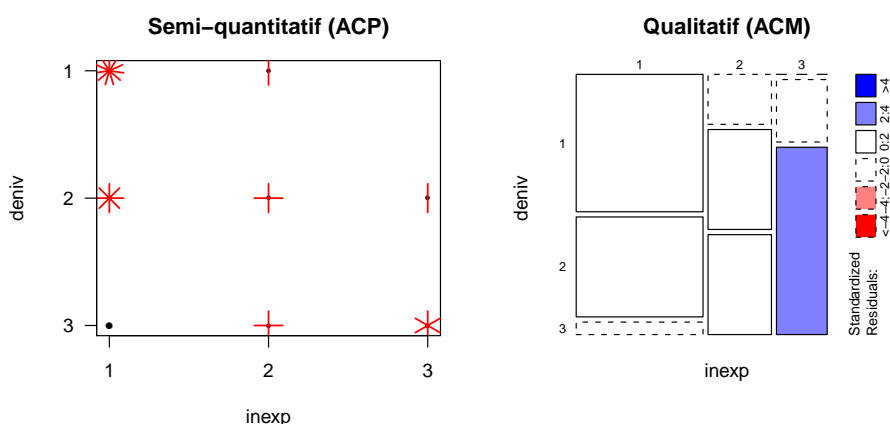


On a donc un gradient entre les régions favorables et défavorables à l'éradication de l'ours brun. Pour bien comprendre la nature des données, on croise la variable 'forêts inexploitées' `inexp` qui est bien corrélée avec la variable 'dénivelé' `deniv`. Du point de vue semi-quantitatif adopté ici, nous avons donc un nuage de points avec beaucoup de superpositions parce qu'il n'y a que peu de valeurs possibles pour les variables. On peut aussi envisager un point de vue complètement qualitatif avec la table de contingence correspondant au croisement de ces deux modalités :

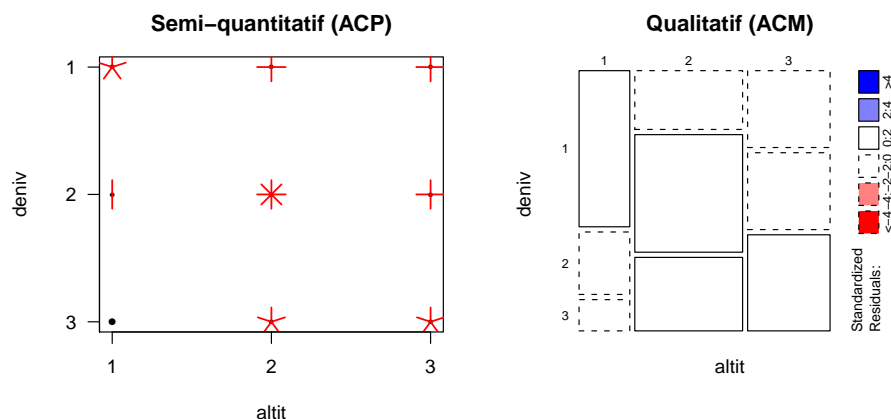
```
table(ours[, c("inexp", "deniv")])
      deniv
inexp  1  2  3
  1  11  8  1
  2   2  4  4
  3   0  2  6
```

La représentation graphique suivante permet de comparer ces points de vue :

```
par(mfrow = c(1, 2), mar = c(5, 4, 3, 1))
sunflowerplot(qtt[, "inexp"], -qtt[, "deniv"], ylab = "deniv", xlab = "inexp",
  main = "Semi-quantitatif (ACP)", xaxt = "n", yaxt = "n")
axis(1, at = 1:3)
axis(2, labels = 1:4, at = -(1:4), las = 1)
mosaicplot(table(ours[, c("inexp", "deniv")]), shade = T, main = "Qualitatif (ACM)",
  las = 1)
```

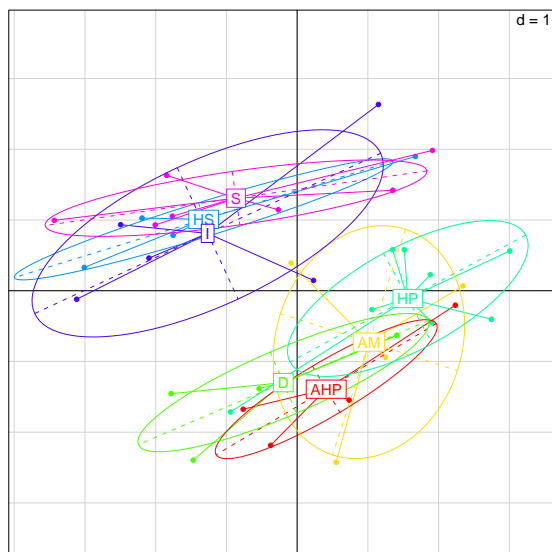


Faire la même représentation pour la variable `altit` pour comprendre en quoi elle diffère des autres :



Nous ne pouvons pas intégrer directement la variable `depart` dans l'ACP parce que c'est d'une part une variable de nature très différente - qui ne décrit pas le milieu - et, d'autre part, une variable qualitative non ordonnée, mais rien ne nous empêche de l'utiliser à titre de variable illustrative.

```
s.class(acp$li, ours$depart, col = rainbow(7))
```



On trouve une opposition entre les départements du nord plus favorables à l'ours que ceux du sud.

4 Une ACM pour continuer

4.1 Introduction

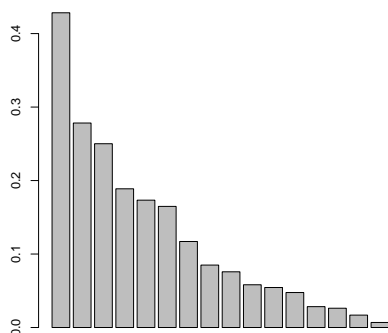
Une variable qualitative ou variable nominale ou facteur (`factor`) est une mesure qui prend ses valeurs dans un ensemble d'items ou modalités ou valeurs ou niveaux (`levels`). La variable `deniv` peut, par exemple, prendre des valeurs dans 1, 2, et 3. Nous allons oublier ici le caractère ordonné de nos variables

pour les traiter comme des variables purement nominales. Une correspondance simple est la présence de deux modalités de deux facteurs différents chez le même individu. Par exemple, comme nous l'avons vu dans le paragraphe précédent, il y a un nombre anormalement grand de régions qui ont à la fois la modalité 3 de la variable `deniv` et la modalité 3 de la variable `inexp`.

Autrement dit, les régions qui présentent d'importantes différences de dénivelée sont aussi des régions difficiles à exploiter par l'homme. Une correspondance multiple est la présence simultanée de m modalités de p facteurs différents chez le même individu. La description du mode d'assemblage des modalités chez les individus relève de l'analyse des correspondances simples (2 facteurs) ou multiples (plus de 2 facteurs). Ici nous avons 8 facteurs environnementaux, c'est donc l'occasion d'utiliser l'analyse des correspondances multiples.

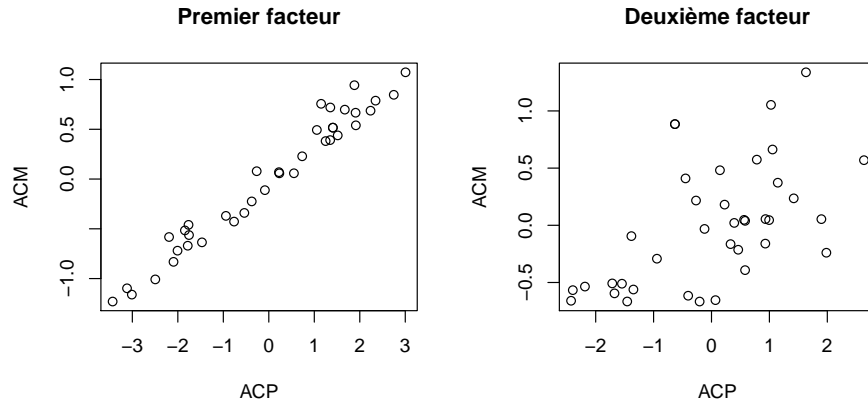
4.2 Mise en oeuvre

```
acm <- dudi.acm(ours[, 1:8], scan = FALSE)
barplot(acm$eig)
```



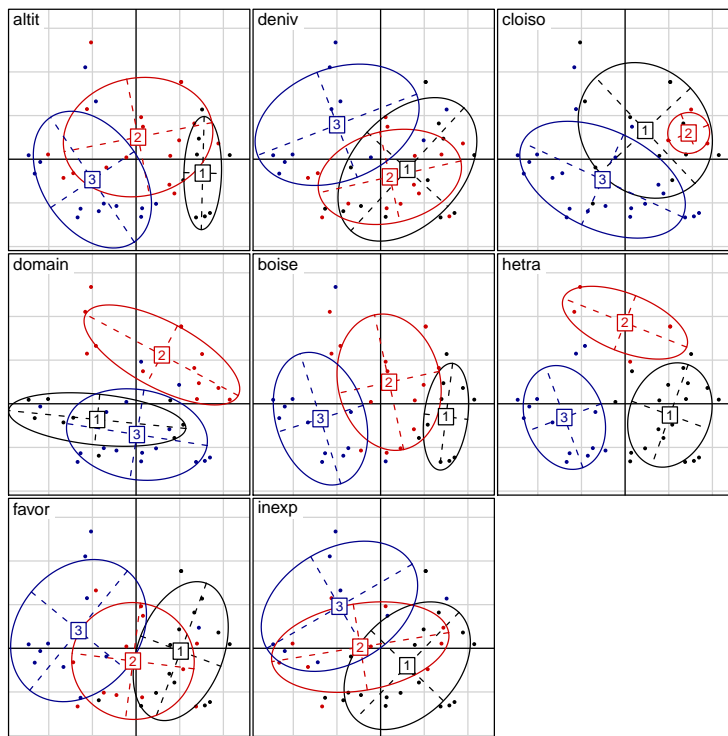
Notez que la décroissance des valeurs propres est bien moins rapide que dans le cas de l'ACP. Il est plus difficile d'énoncer un critère relatif au nombre de facteurs à conserver. On décide de ne considérer que les deux premiers facteurs par la suite. Comparons les coordonnées des individus sur le premier et le deuxième facteurs entre les deux analyses :

```
par(mfrow = c(1, 2))
plot(acp$li[, 1], acm$li[, 1], xlab = "ACP", ylab = "ACM", main = "Premier facteur")
plot(acp$li[, 2], acm$li[, 2], xlab = "ACP", ylab = "ACM", main = "Deuxième facteur")
```



Les résultats sont très comparables : l'ACM retrouve les mêmes structures que celles fournies par l'ACP, ce qui est rassurant car on n'aurait pas aimé passer à côté d'une structure aussi forte. Examinons le premier plan factoriel :

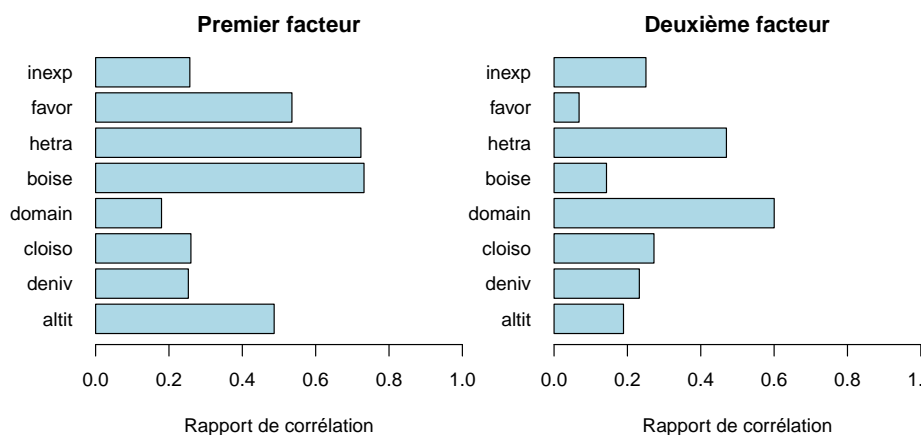
```
scatter(acm, col = rep(c("black", "red3", "darkblue"), 2))
```



Le même plan factoriel est répété autant de fois qu'il y a de variables dans le tableau de départ, ici 8. Sur chaque plan il y a 38 points, correspondant aux 38 régions analysées. Pour faciliter l'interprétation on représente, variable par variable, la modalité prise par chaque individu avec un code couleur et une ellipse résumant la dispersion des points. On voit par exemple que pour la variable `altit` le premier facteur oppose les régions ayant la modalité 1 (faible altitude) aux régions ayant la modalité 3 (altitude élevée). Examiner soigneuse-

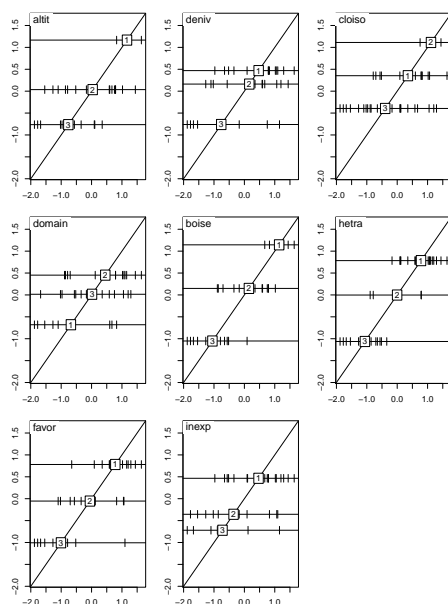
ment le premier plan factoriel variable par variable. Retrouve-t-on les résultats de l'ACP? Que pouvez-vous dire de la variable `domain`? L'objectif de l'ACM est d'obtenir des scores numériques des individus qui maximisent les pourcentages de variance expliquée, en moyenne, pour toutes les variables qualitatives (rapport de corrélation). Il est donc intéressant d'examiner ces rapports de corrélation entre les variables de départ et les facteurs interprétés. Faites le lien avec les plans factoriels.

```
par(mfrow = c(1, 2), mar = c(5, 4, 2, 0))
barplot(acm$scr[, 1], horiz = TRUE, xlim = c(0, 1), names.arg = colnames(ours[1:8]),
       las = 1, main = "Premier facteur", col = "lightblue", xlab = "Rapport de corrélation")
barplot(acm$scr[, 2], horiz = TRUE, xlim = c(0, 1), names.arg = colnames(ours[1:8]),
       las = 1, main = "Deuxième facteur", col = "lightblue", xlab = "Rapport de corrélation")
```



Une autre aide à l'interprétation des facteurs est donnée par la fonction `score()`, qui est une version univariée des plans factoriels précédents. Pour chaque variable, les individus sont positionnés sur l'axe des abscisses par leur score sur l'axe factoriel considéré et sur l'axe des ordonnées par le score de la modalité qu'ils portent. Le score d'une modalité n'est rien d'autre que la moyenne des scores des individus portant cette modalité, ce qui est mis en évidence par la première bissectrice dans les graphiques suivants.

```
score(acm, xax = 1)
```



En étudiant également le résultat de `score(acm, xax = 2)`, faites le lien avec les plans factoriels.

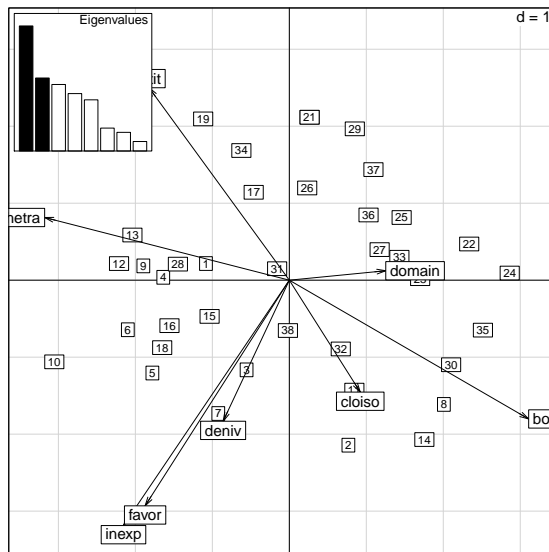
4.3 Une différence importante en l'ACP et l'ACM

La grande similitude des résultats ici entre l'ACP et l'ACM ne doit pas masquer une différence fondamentale entre les deux méthodes : dans l'ACM, l'ordre des modalités est perdu. Comme cela correspond à une perte d'information, on pourrait penser que c'est un désavantage de l'ACM. Ce n'est pas forcément le cas, et nous allons l'illustrer avec une petite expérience où nous mélangeons sauvagement les modalités des données de départ.

```
melanger <- function(facteur) {
  new <- factor(facteur)
  levels(new) <- sample(levels(new))
  return(new)
}
ours.rnd <- as.data.frame(apply(ours[, 1:8], 2, melanger))
summary(ours[, 1:8])
altit  deriv  cloiso domain boise  hetra  favor  inexp
1: 8    1:13  1:12  1: 9    1:10  1:19  1:15  1:20
2:17   2:14  2: 4    2:13   2:15  2: 5   2:12  2:10
3:13   3:11  3:22  3:16   3:13  3:14  3:11  3: 8
summary(ours.rnd)
altit  deriv  cloiso domain boise  hetra  favor  inexp
1: 8    1:13  1:12  1: 9    1:13  1:19  1:15  1:20
2:13   2:11  2:22  2:13   2:15  2: 5   2:11  2:10
3:17   3:14  3: 4    3:16   3:10  3:14  3:12  3: 8
```

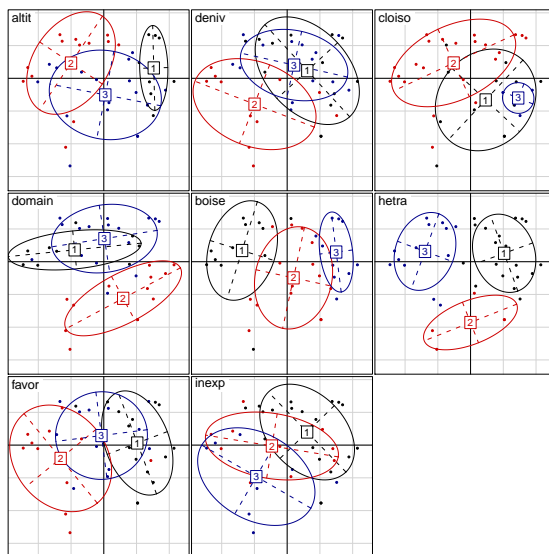
Que donne l'ACP sur ces données mélangées ?

```
scatter(dudi.pca(apply(ours.rnd[, 1:8], 2, as.numeric), scan = FALSE))
```



Nous avons complètement détruit la structure initiale des données, il n'est pas sans conséquence de détruire l'ordre d'une variable ordonnée. Que donne l'ACM sur les données mélangées ?

```
scatter(dudi.acm(ours.rnd, scan = FALSE), col = rep(c("black", "red3",
"darkblue"), 2))
```

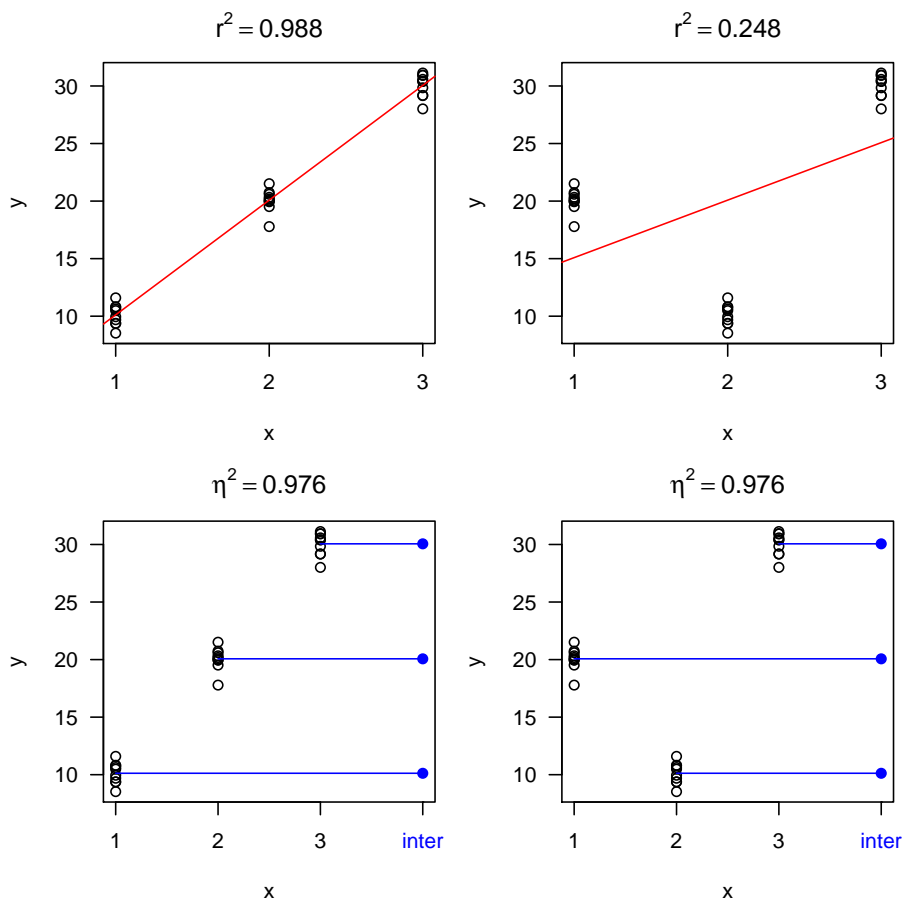


Aux opérations de symétrie près, on retrouve exactement les mêmes résultats que pour l'ACM de départ. On a bien entendu perdu la signification des modalités, mais la structure du nuage de point est inchangée. Il est sans conséquence (autre que sémantique!) de changer les noms d'une variable nominale. Si la relation d'ordre entre les modalités des variables étudiées n'est pas absolue, comme c'est ici le cas avec la variable **domain**, il peut être intéressant de privilégier l'ACM à l'ACP pour ne pas tenir compte d'un ordre que l'on sait ne pas être absolu. Les graphiques suivants avec des données simulées devraient vous aider à comprendre cette différence entre l'ACP et l'ACM.

```

set.seed(1)
x <- rep(1:3, 10)
x2 <- rep(c(2, 1, 3), 10)
y <- rnorm(30, c(10, 20, 30))
pltr2 <- function(x, y) {
  r2 <- cor(x, y)^2
  plot(x, y, xaxt = "n", las = 1, main = bquote(r^2 == .(round(r2,
    3))))
  axis(1, at = 1:3)
  abline(lm(y ~ x), col = "red")
}
plteta2 <- function(x, y) {
  x.q <- factor(x)
  var.n <- function(x) sum((x - mean(x))^2)/length(x)
  vartot <- var.n(y)
  moys <- by(y, x.q, mean)
  varinter <- var.n(moys)
  e2 <- (varinter/vartot)^2
  plot(x, y, xaxt = "n", las = 1, main = bquote(eta^2 == .(round(e2,
    3))), xlim = c(min(x), max(x) + 1))
  axis(1, at = 1:3)
  axis(1, "inter", at = 4, col.axis = "blue")
  points(rep(4, 3), moys, pch = 19, col = "blue")
  segments(1:3, moys, rep(4, 3), moys, col = "blue")
}
par(mfrow = c(2, 2), mar = c(4, 4, 3, 1))
pltr2(x, y)
pltr2(x2, y)
plteta2(x, y)
plteta2(x2, y)

```



Références

- [1] G. Erome. *L'ours brun dans les Alpes françaises. Historique de sa disparition*. Centre Ornithologique Rhône-Alpes, Villeurbanne, 1989.